

# CLASS-IT: Conversational and Lecture-Aligned Small-Scale Instruction Tuning for BabyLMs

Luca Capone<sup>1,\*†</sup> and Alessandro Bondielli<sup>1,2†</sup> and Alessandro Lenci<sup>1†</sup>

<sup>1</sup>CoLing Lab, Department of Philology, Literature and Linguistics, University of Pisa

<sup>2</sup>Department of Computer Science, University of Pisa

luca.capone@fileli.unipi.it, {alessandro.bondielli, alessandro.lenci}@unipi.it

## Abstract

This work investigates whether small-scale LMs can benefit from instruction tuning. We compare conversational and question-answering instruction tuning datasets, applied either in a merged or sequential curriculum, using decoder-only models with 100M and 140M parameters. Evaluation spans both fine-tuning (SuperGLUE) and zero-shot (BLiMP, EWoK, WUGs, entity tracking, and psycholinguistic correlation) settings. Results show that instruction tuning yields small but consistent gains in fine-tuning scenarios, with sequential curricula outperforming merged data; however, improvements do not consistently transfer to zero-shot tasks, suggesting a trade-off between interaction-focused adaptation and broad linguistic generalization. These results highlight both the potential and the constraints of adapting human-inspired learning strategies to low-resource LMs, and point toward hybrid, curriculum-based approaches for enhancing generalization under ecological training limits.

## 1 Introduction

The role of input data vis-à-vis innate biases has long dominated the debate on language acquisition. This is exemplified by arguments such as the poverty of the stimulus and the language of thought hypothesis (Chomsky, 1980; Fodor, 1975), which have emphasized the need for innate constraints governing the process of acquiring productive linguistic generalizations. In contrast, data-driven learning has always been a central tenet of connectionist theory, arguing that, given sufficient training, a large enough model can reproduce any regular behavioral pattern (Smolensky, 1988). One of the defining features of LMs is that performance relies on the training process. The development of

model abilities clearly reflects learning, although the precise nature of this learning is not yet well understood. It remains uncertain whether abilities (or at least some of them) are truly emergent (Wei et al., 2022), or whether this impression is an artifact of measurement, with capabilities in fact increasing more gradually (Schaeffer et al., 2023). Moreover, the type and order of training data can influence a model’s ability to perform specific tasks (Soviany et al., 2022). Finally, particular training regimes, such as instruction tuning or reinforcement learning with human feedback (RLHF), can significantly enhance a model’s capacity for user interaction, as well as its logical, inferential, and reasoning abilities. Despite relying on radically different mechanisms, LMs and humans share several key properties of learning: both improve with training over time, both are sensitive to the quality of instruction, and both benefit from interactive, feedback-driven training. These parallels suggest that current LMs approximate some aspects of human-like learning. However, the scale of resources required (both in terms of data and computation) remains orders of magnitude greater than what is needed for human learning, especially in children (Frank, 2023). Among these shared features, this paper focuses on **interaction**, a core component of human learning, particularly in childhood. We investigate whether an LM trained on ecologically valid input, comparable in scale to the linguistic exposure of a 10-year-old child, can benefit significantly from targeted instruction tuning. Specifically, we compare two types of instruction tuning datasets: one centered on conversational interactions and the other focused on question-answering tasks. The main research questions addressed in this study are:

- Can a BabyLM benefit from instruction tuning?
- Given the limited pre-training typical of BabyLMs, which type of instruction data is

\*Corresponding author

†For the specific purposes of Italian Academy, Luca Capone is responsible for Sections 2, 3 and 4, Alessandro Bondielli is responsible for sections 5 and 6, Alessandro Lenci is responsible for sections 1 and 7.

more effective: conversational or open-ended question-answering?

- Does a curriculum learning approach to instruction tuning provide significant benefits?

This paper is organized as follows. Section 2 reviews related work. Section 3 describes the datasets used for pre-training and instruction tuning. Section 4 presents the model architectures and details the training procedures, while Sections 5 and 6 present and analyze the results on the BabyLM Challenge tasks. Finally, Section 7 summarizes our findings and outlines directions for future research.

## 2 Related Works

While early language learning in children is often portrayed as remarkably precocious (McCormack and Hoerl, 2005; Gopnik, 2011; Dündar-Coecke et al., 2020), linguistic and psychological studies suggest that this view must be qualified. Many scholars acknowledge children’s early communicative abilities, but argue that these are constrained to specific tasks and contexts, and do not necessarily reflect a fully developed understanding of language. For instance, although the intersubjective (i.e., social and communicative) function of linguistic signs becomes evident in children from an early age, their perspectival function, the ability to conceptualize experiences from multiple viewpoints, emerges more gradually (Vygotsky, 1987; Piaget, 2002; Tomasello, 2009).

Drawing on developmental psycholinguistic evidence (Berman and Slobin, 2013; Peterson and McCabe, 1987), Tomasello (2003) observed that many children up to the age of nine, despite producing fluent, age-appropriate speech, struggle to use sophisticated conjunctions (such as *because*, *indeed*, *although*, etc.) when required to do so. These conjunctions involve representing events from a logical-causal or antithetical perspective, which can pose significant challenges. At this stage, *and* remains the most frequently used connective, functioning in an undifferentiated way to express a wide range of semantic relations, even after more specific connectives have begun to appear in a child’s speech. Similar limitations occur with other complex constructions: comprehension and voluntary use often do not match the apparent fluency of spontaneous speech. Berman and Slobin (2013) document the difficulties children face with narrative discourse, sometimes even up to age nine,

when asked to describe a story depicted in a sequence of images. Children frequently struggle to produce coherent narratives that clearly indicate a beginning, progression, and conclusion. Tomasello (2003) attributes these challenges to the *plurifunctionality* of complex constructions, arguing that mastery of the perspectives they encode develops gradually over the course of the school years.

Building on this body of research, the present study investigates whether and to what extent interactive instruction can enhance the training of a BabyLMs. In particular, it examines whether formal, instruction-like input provides greater benefits than conversational data for fine-tuning LMs trained on limited, child-comparable linguistic exposure. To our knowledge, the two main attempts to interactively train BabyLMs using more pedagogically structured data are Baby’s CoThought (Zhang et al., 2023) and Baby Stories (Zhao et al., 2023). However, both differ from the approach proposed in this work, albeit for different reasons. Zhang et al. (2023) build an educational dataset based on the BabyLM Challenge trainset, using GPT-3.5-Turbo. However, the dataset is used to train an encoder-only model through masked language modeling. Zhao et al. (2023), on the other hand, preserves an interactive setup by fine-tuning a decoder model using proximal policy optimization. Nonetheless, this training technique departs from the type of formal instruction we aim to address, as the model is simply optimized to prefer certain generations based on a reward model. In contrast, the instruction tuning proposed in this work more closely resembles the structured, formal education typically provided to children in school settings. The present work instead adopts an instruction fine-tuning approach, where models are explicitly trained to respond to questions about specific topics and to provide appropriate answers within conversational contexts.

## 3 Dataset

The dataset used for model pre-training is a curated subset of the data provided by the task organizers, which amounts to approximately 91 million words. The instruction tuning dataset includes processed Switchboard transcripts and augmented Simple Wikipedia texts, enhanced using the LLaMA-3.2-3B-Instruct model (Dubey et al., 2024).

3.1 Pretraining Dataset

The data supplied by the organizers (approximately 100 million words) underwent standard preprocessing. Special characters were removed, and all entries containing two words or fewer were discarded. Additional processing was applied to the Switchboard corpus, utterances from the same speaker were concatenated when they occurred in sequence, following standard dialogue normalization practices. Roughly 75 million words—drawn from CHILDES, Gutenberg, BNC and OpenSubtitles—were used exclusively for pre-training. An additional 16 million words (from Switchboard and Simple Wikipedia) overlap with the instruction tuning dataset, bringing the total pre-training corpus to approximately 91 million words.

3.2 Instruction-Tuning Dataset

The instruction tuning dataset consists of two sections: a **conversational component** based on the Switchboard corpus and an **instructional component** based on Simple Wikipedia. For the conversational section, the Switchboard data were adapted to meet the requirements of instruction tuning training task. Consecutive utterances from the same speaker were merged to ensure a consistent alternation between speakers’ turns (e.g., A, B, A, B). The dialogues were then segmented into prompt–reply pairs using a sliding window approach with the following schema: (A1, B1), (B1, A2), (A2, B2). The resulting dataset contains 38,802 items and approximately 1.3 million words (excluding prompt–reply duplicates). For the instructional section, Simple Wikipedia data were augmented using LLaMA-3.2-3B-Instruct (Dubey et al., 2024). For each article text, three question–answer pairs were generated using structured generation with *outlines*<sup>1</sup> and the following prompt:

Based on the following text, generate 3 questions and detailed, informative answers. Each answer should be easy for a young person to understand and at least 2–3 sentences long. Explain things in simple language, with clear and friendly sentences. Avoid short or vague replies and give enough detail so a kid can learn something new.

The generated data significantly exceeds the

<sup>1</sup><https://dottxt-ai.github.io/outlines/latest/#acknowledgements>

| Hyperparameter       | llama140M   | llama100M   |
|----------------------|-------------|-------------|
| Vocab size           | 32,000      | 16,384      |
| Max length           | 6,144       | 6,000       |
| Hidden size          | 704         | 512         |
| Attention heads      | 11          | 8           |
| Layers               | 12          | 20          |
| Trainable parameters | 140,231,872 | 100,684,288 |

Table 1: Model architectures

| Hyperparameter | Pretrain | Instr. tuning      |
|----------------|----------|--------------------|
| Initial LR     | 2e-4     | 2e-5               |
| Batch size     | 8        | 8                  |
| Maximum epochs | 8        | 10                 |
| LR scheduler   | linear   | cosine w/ restarts |
| Warm-up steps  | 5,000    | 500                |

Table 2: Training parameters

word limit imposed by the challenge. In this work we use only a representative portion of the whole dataset (colinglab/CLASS\_IT). The full dataset will be released in the future following appropriate validation. The subset used in this study contains 8.7 million words, keeping the total—along with the 91 million pre-training words (which already include Switchboard and Simple Wikipedia texts)—within the 100-million-word limit. The augmented Simple Wikipedia dataset includes 97,697 items, totalling 18 million words (Figure 1).

4 Models and training

We trained two models (Table 1), both based on decoder-only, LLaMA-style architectures with large maximum sequence lengths to accommodate the long texts present in the instruction tuning dataset. The first model has 140 million parameters, featuring a larger hidden size and vocabulary size. Following its training, we developed a second model with approximately 100 million parameters, using a reduced hidden size and a vocabulary size comparable to baseline models.

Both models followed the same pre-training procedure. The tokenizer was trained on the entire available corpus before pre-training began. Models were then pre-trained for 8 epochs using the parameters in Table 2, processing a total of approximately 728 million words.

Instruction tuning use a different set of hyperpa-

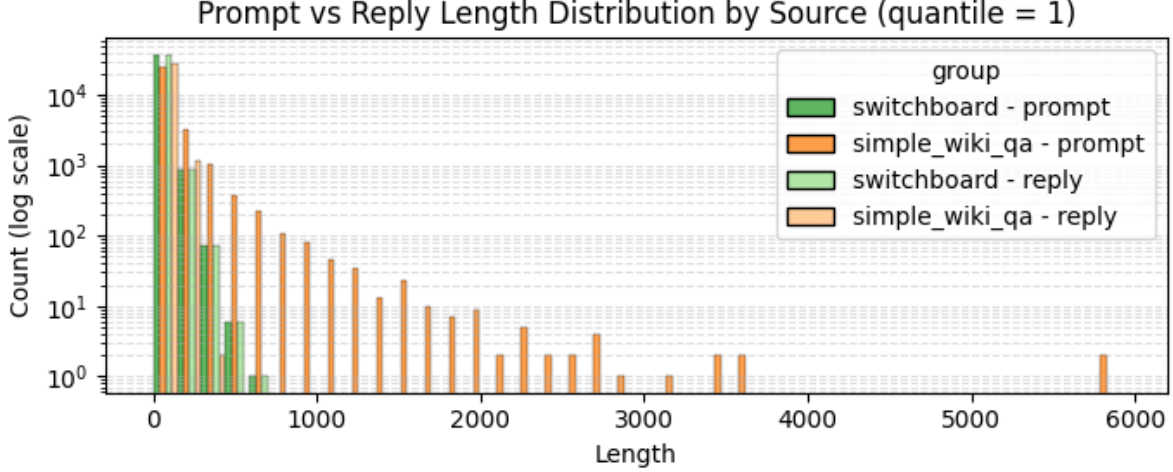


Figure 1

rameters from pre-training, but the same instruction tuning configuration was applied to both models (see Table 2). Each model was fine-tuned for 10 epochs, processing an additional 180 million words. In total, each model processed around 908 million words across both pre-training and instruction tuning. All datasets were split 90/10 into training and validation sets, with only the training portion contributing to parameter updates. Consequently, roughly 90% of the 908 million words (about 817 million) directly influenced model weight updates. We used the same token-level cross-entropy loss used for pre-training. However, for instruction tuning, we compute the loss only on target tokens, e.g. the answer tokens in a question-answer data point.

We adopted two strategies for instruction tuning: **merged** and **sequential**. In the merged strategy, augmented Simple Wikipedia data were shuffled together with Switchboard data, mixing conversational and instructional items. This produced the `it_merged` models (see Figure 2). In the sequential strategy, the two datasets were used in succession, resulting in two variants: `it_switch_wiki` and `it_wiki_switch`, depending on the order in which the pre-trained model was exposed to the instruction tuning datasets. This approach was designed to test whether keeping the tasks separate—and whether the order of exposure—provides measurable benefits to model performance.

## 5 Evaluation and Results

To evaluate our models, we used the official data provided by the challenge organizers. The evaluation is distinguished between a fine-tuning evalua-

tion and a zero-shot evaluation.

**Fine-Tuning Evaluation.** In the fine-tuning evaluation, the models are fine-tuned and evaluated in the (Super)GLUE (Sarlin et al., 2020) tasks. We leave all the default parameters unchanged during training on each task. Note that the fine-tuning dataset is composed of a randomly sampled 10k portion of the original training set for the task. Models are evaluated on the test set. Figure 2 shows the result of our models (both pre-trained and instruction-tuned) and the baselines (`bl_gpt2-100M`, `bl_gptbertmixed-100M`, `bl_simpo`, shown in blue).

We observe that our models are generally competitive with the baselines, albeit surpassing them only for some configurations in the WSC task. As for the model size, the 140 million parameter models are generally better than the 100 million ones. However, the only tasks where the difference is noticeable are QQP and MNLI. Here, the 100 million models are markedly worse than their 140 million siblings, and both are markedly worse than the baselines. As for the instruction fine-tuning, it seems relatively beneficial. We see in fact that in all cases there is at least an instruction-tuned model better than the pre-trained one. However, differences are small and inconsistent across tasks, that is, there is no instruction tuning configuration that systematically leads to better results.

Since we have multiple tasks and models, we needed a way to compare performance globally. To achieve this, we standardized the results by computing z-scores for each model on each task, which express how many standard deviations above or



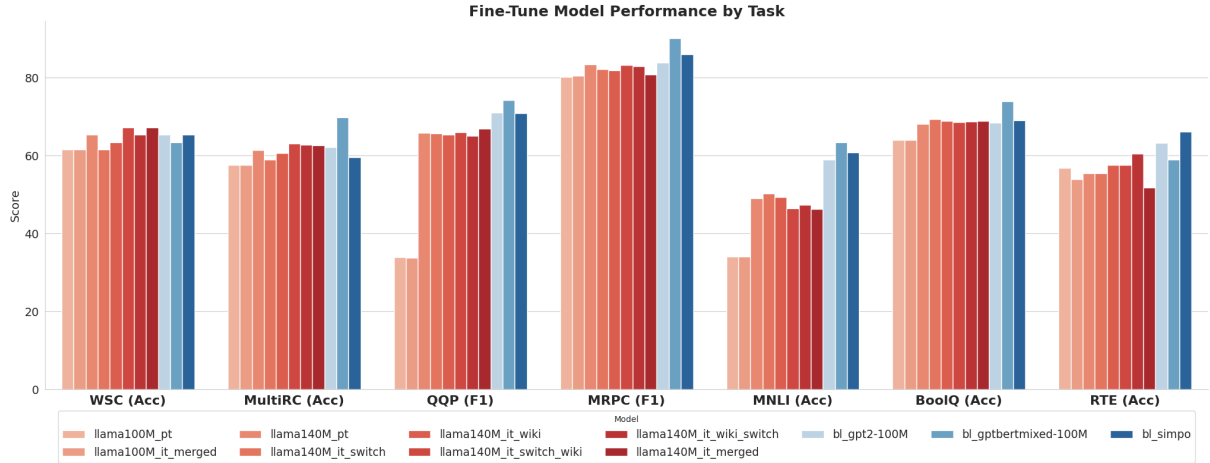


Figure 2: Results of fine-tuned models on (Super)Glue tasks.

below the task mean a model’s score lies. We then averaged these z-scores across tasks to obtain a single global index per model. This index reflects the overall relative standing of a model compared to others, rather than absolute task performance, and allows fair comparison across heterogeneous metrics. Specifically, we plot them including median, Inter-Quartile Ranges (IQR), and outliers. Results are in Figure 3.

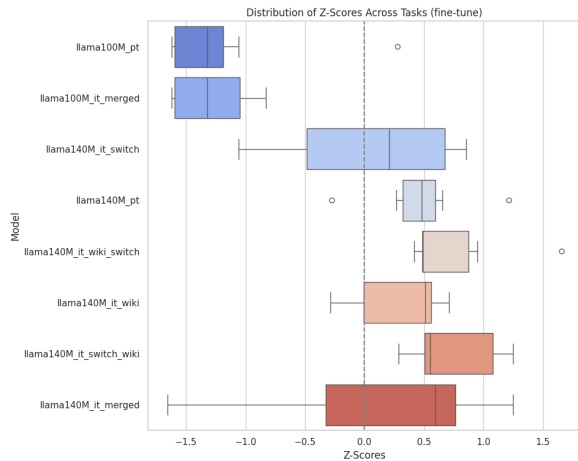


Figure 3: Median, Inter-Quartile Ranges (IQR), and outliers for z-scores of each model in the **fine-tuning** evaluation.

We observe that the 100 million models, both pre-trained and instruction-tuned, have negative z-score medians, while all the 140 million variants are on the positive side of the plot, showcasing that differences between the smaller and larger models appear to be significant. Regarding the differences between pre-training only and instruction tuning, we notice some interesting aspects. The model with the highest median is the merged instruction-

tuned variant trained on a mixture of the datasets. However, both models trained on the two dataset sequentially, regardless of the order, have a very similar median score, but a much smaller IQR, and are the only two models with all z-scores above zero. Variants trained only on one of the datasets perform worse, and on par with the pre-trained only model, with the one trained just on Switchboard performing worst.

**Zero-shot Evaluation.** In the zero-shot scenario, models are evaluated using log-probabilities of sequences and/or words to obtain either model predictions or compute correlations with human data. The zero-shot evaluation is conducted on the following datasets: BLiMP (Warstadt et al., 2020) and EWoK (Ivanova et al., 2024) are standard minimal pairs datasets that test linguistic and world knowledge of LLMs and were included also in previous years’ evaluations; a WUGs task designed to understand abilities in adjective nominalization (Hofmann et al., 2025); an entity tracking task on data from (Kim and Schuster, 2023); a correlation evaluation where cloze probability, predictability ratings, and computational estimates are compared against EEG and human reading time data (de Varda et al., 2024).

Results are reported in Figure 4. For the accuracy-based tasks, we do not observe striking differences between pre-trained and instruction-tuned models, similarly to what seen in the fine-tuning evaluation. However, here we also do not observe large differences also between the 100 million and 140 million variants. Our models seem to vastly outperform baselines on the WUGs task, but are worse on the Entity Tracking task, on which all

models including baselines seem to struggle. As for the Change in  $R^2$  based tasks, we observe some surprising results: The 100 million model variants are vastly superior to both the 140 million models and the baselines, which score almost zero with the exception of the GPT-BERT mixed model. We compute z-scores distribution also in this case, and report them in Figure 5.

For the zero-shot evaluation the z-score distribution is radically different. No model has all z-scores above zero, and only two of them has a median z-score above zero. The two 100 million variants are among the best performing models, albeit this could be attributed to the vast differences between them and all the other models on the  $R^2$ -based tasks. The best performing model is an instruction-tuned variant, specifically the one trained only on Simple Wikipedia. However, no clear trend in favour or against instruction tuning emerge from the plot.

In order to further examine the performances on a broader level, we also plot the z-score distribution including both zero-shot and fine-tuning evaluations. Results are shown in Figure 6. It highlights the fact that, overall, the larger models seem to perform better. As for the impact of instruction tuning, we can highlight three aspects. First, we see that the best overall model is an instruction-tuned one. However, we cannot extrapolate a clear trend in favour of instruction tuning. Second, we observe that tuning the model sequentially on different datasets is consistently better than doing so on a mixture of the datasets. The order of the instruction tuning task seems less relevant, albeit we see that tuning first on conversational data (Switchboard) and then on question answering (Simple Wiki) seem to yield better results. This however may be affected by the difference in size between the datasets. In fact, we see that the model trained only on question answering performs better than the one trained subsequently on conversations.

## 6 Discussion

Our experiments provide some interesting insights about small-scale instruction tuning models trained on ecological amounts of data.

First, we see that instruction tuning appears to be somewhat beneficial, especially if the model is further fine-tuned on specific tasks; the same improvement are not as apparent on the zero-shot evaluation. We can hypothesize that the instruction

tuning stage varies the models' internal distribution to a higher degree, especially at this scale, thus affecting the performances on zero-shot tasks, where the encoding of grammar rules (BLiMP, WUGs) or specific facts (EWoK) is more relevant than conversational and/or generative performances, which are not tested here. The instruction tuned models may be biased to learn to solve a specific task, in our case following a conversation or answering factual questions, thus losing their generalization abilities on just language. This latter aspect is quite interesting in the context of BabyLMs, as larger models have been shown to not suffer from similar issues (Miliiani et al., 2025). Further evidence for this effect can be seen in the fact that instruction-tuned models have seen more data than pre-trained ones, yet do not consistently outperform them.

Second, we observe that among our models, smaller ones are consistently better than larger ones at correlating with human data, with the pre-trained model being slightly better; only one baseline model, with a different architecture but of the same size as ours, achieve comparable results. This is in line with previous literature where smaller models often correlate better with human psychometric data (De Varda and Marelli, 2023; Oh and Schuler, 2023).

Finally, we see that all our models achieve relative consistent performance on a single-task basis, while being very inconsistent across tasks and evaluation methods; the same happens with baselines. This suggests that the constraints posed by the challenge itself, namely the amount of data and training compute allowed for a training run, may limit the generalization capability of decoder-only style models without additional modifications.

## 7 Conclusion and future work

This study examined whether BabyLM-scale models (trained on ecologically realistic amounts of linguistic input) can benefit from instruction tuning, and how different forms and orders of data affect their performance. Our findings show that instruction tuning yields modest but measurable gains in fine-tuning scenarios, particularly when conversational and question-answering datasets are presented sequentially rather than merged. However, these benefits do not translate consistently to zero-shot evaluations, suggesting that, at this scale, instruction tuning may bias models toward narrow interactional behaviors at the expense of broader

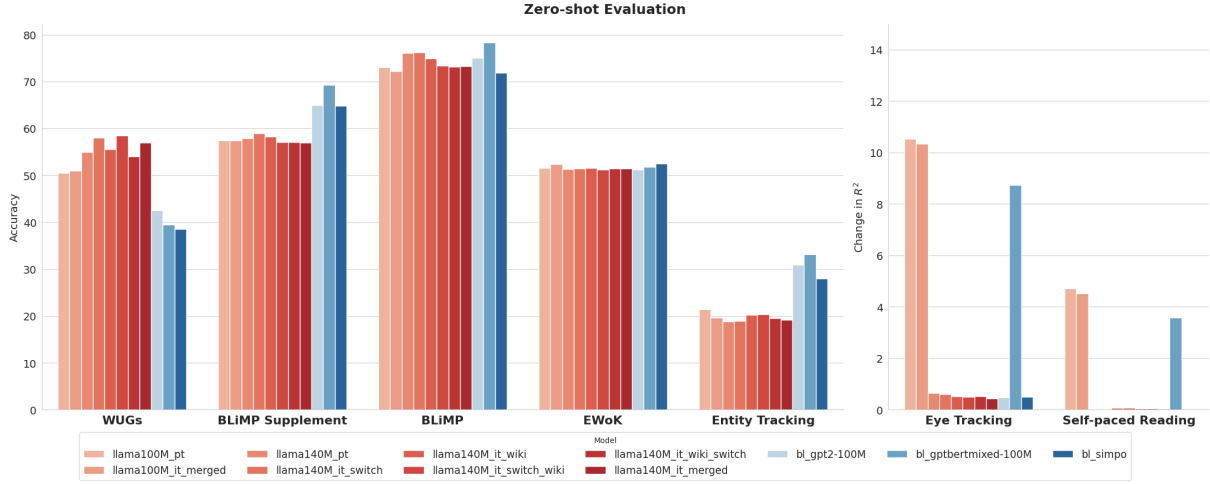


Figure 4: Results of the zero-shot evaluation. Tasks measured with accuracy are reported in the left bar chart; tasks measured with change in  $R^2$  are reported in the bar chart on the right.

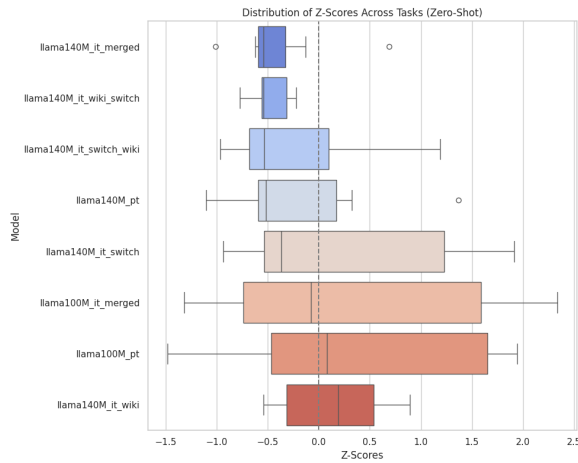


Figure 5: Median, Inter-Quartile Ranges (IQR), and outliers for z-scores of each model in the **zero-shot** evaluation.

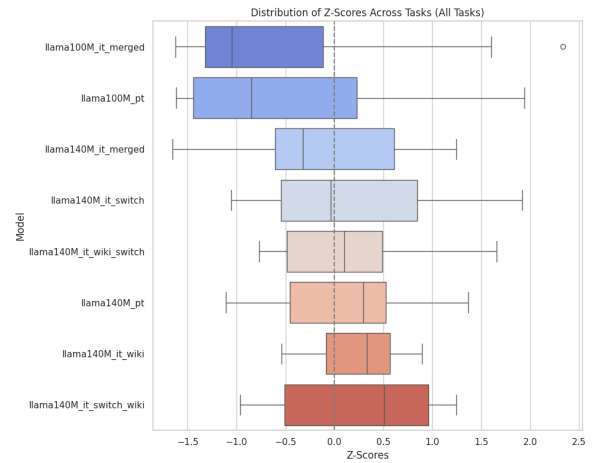


Figure 6: Median, Inter-Quartile Ranges (IQR), and outliers for z-scores of each model including both zero-shot and fine-tuning evaluations.

linguistic generalization.

A further limitation lies in how models are evaluated in the Challenge. In fact, in the fine-tuning evaluation most tasks are actually classification tasks, on which masked LMs may prove more reliable; in the zero-shot task, the vast majority of evaluations are conducted using log-likelihood as proxy for model choices. While this choice is valid in the context of the challenge, to accommodate the largest possible number of architectures and simplify the evaluation process, we can argue that model performances, especially when considering conversational instruction tuning, may be undermined by the evaluation criteria. Moreover, the chosen conversational portion of the instruction tuning dataset may limit the performances of the

model: while the Switchboard corpus offers a structured and well-annotated resource, it represents a restricted register of spoken English and lacks much of the contextual diversity found in everyday interaction. More ecologically valid conversational data, spanning a wider range of speakers, settings, and discourse types, would provide a richer foundation for model adaptation and a stronger basis for subsequent instructional fine-tuning, potentially improving both interactive competence and generalization. Notably, smaller models exhibited stronger correlations with human psycholinguistic data, echoing prior observations that reduced capacity can sometimes yield representations more aligned with human processing patterns. Overall, the results highlight both the promise and the limi-

tations of adapting human-inspired learning strategies to small-scale LMs: interaction helps, but the gains are context-dependent, and generalization remains challenging under strict data and compute constraints. Future work should explore hybrid approaches that combine instruction tuning with targeted multi-task or curriculum learning, investigate architectures better suited for low-resource generalization, and extend the evaluation to interactive and communicative benchmarks that more directly reflect the ecological learning conditions motivating the BabyLM challenge.

## Limitations

Our instruction tuning experiments are constrained by the relatively small size of the instruction tuning datasets compared to pre-training corpus, which may have reduced the impact of instruction-specific learning. A different allocation (using more instruction tuning data and proportionally less pre-training data) might yield stronger effects. Moreover, the balance between question–answering and conversational data is imperfect, with the latter under-represented, potentially biasing results toward factual over interactive skills. Finally, the Simple Wikipedia augmentation process was only partially validated, and higher-quality or more diverse instructional sources could improve both robustness and generalization.

## Acknowledgments

We acknowledge financial support under the PRIN 2022 Project Title "Computational and linguistic benchmarks for the study of verb argument structure" – CUP I53D23004050006 - Grant Assignment Decree No. 1016 adopted on 07/07/2023 by the Italian Ministry of University and Research (MUR). This work was also supported under the PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013—"FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme"

## References

- Ruth A Berman and Dan Isaac Slobin. 2013. *Relating events in narrative: A crosslinguistic developmental study*. Psychology Press.
- Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

- Andrea De Varda and Marco Marelli. 2023. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149.
- Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 56(5):5190–5213.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *CoRR*.
- Selma Dünder-Coecke, Andrew Tolmie, and Anne Schlottmann. 2020. Children’s reasoning about continuous causal processes: The role of verbal and non-verbal ability. *British Journal of Educational Psychology*, 90(2):364–381.
- Jerry Fodor. 1975. The language of thought (new york: Thomas crowell).(1987a). *Psychosemantics: the Problem of Meaning in the Philosophy of Mind*.
- Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992.
- Alison Gopnik. 2011. The theory theory 2.0: probabilistic models and cognitive development. *Child Development Perspectives*, 5(3):161–163.
- Valentin Hofmann, Leonie Weissweiler, David R Mortensen, Hinrich Schütze, and Janet B Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19):e2423232122.
- Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas Hikaru Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, and 1 others. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *CoRR*.
- Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Teresa McCormack and Christoph Hoerl. 2005. Children’s reasoning about the causal significance of the temporal order of events. *Developmental Psychology*, 41(1):54.
- Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. 2025. Explica:



- Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.
- Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.
- Carole Peterson and Allyssa McCabe. 1987. The connective ‘and’: Do older children use it less as they learn other connectives? *Journal of Child Language*, 14(2):375–381.
- Jean Piaget. 2002. *Judgement and reasoning in the child*. Routledge.
- Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.
- Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581.
- Paul Smolensky. 1988. On the proper treatment of connectionism. *Behavioral and brain sciences*, 11(1):1–23.
- Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.
- Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.
- Michael Tomasello. 2009. *The cultural origins of human cognition*. Harvard university press.
- Lev Semenovich Vygotsky. 1987. *The collected works of LS Vygotsky: Volume 1: Problems of general psychology, including the volume Thinking and Speech*, volume 1. Springer Science & Business Media.
- Samuel R Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english (electronic resources).
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Zheyu Zhang, Han Yang, Bolei Ma, David Rügamer, and Ercong Nie. 2023. Baby’s cothought: Leveraging large language models for enhanced reasoning in compact models. *arXiv e-prints*, pages arXiv–2308.
- Xingmeng Zhao, Tongnian Wang, Sheri Osborn, and Anthony Rios. 2023. Babystories: Can reinforcement learning teach baby language models to write better stories? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 186–197.